A Multiple-Hypothesis Tracking of Multiple Ground Targets from Aerial Video with Dynamic Sensor Control*

Pablo Arambel, Matthew Antone, Constantino Rago, Herbert Landau
ALPHATECH, Inc,
6 New England Executive Park,
Burlington, MA 01803
USA
parambel@alphatech.com

Thomas Strat
DARPA Information Exploitation Office,
3701 N Fairfax Drive,
Arlington, VA 22203
USA

Abstract* – *The goal of the DARPA Video Verification of Identity (VIVID) program is to develop an automated video-based ground targeting system for unmanned aerial vehicles that significantly improves operator combat efficiency and effectiveness while minimizing collateral damage. One of the key components of VIVID is the Multiple Target Tracker (MTT), whose main function is to track many ground targets simultaneously by slewing the video sensor from target to target and zooming in and out as necessary. The MTT comprises three modules: (i) a video processor that performs moving object detection, feature extraction, and site modeling; (ii) a multiple hypothesis tracker that processes extracted video reports (e.g. positions, velocities, features) to generate tracks of currently and previously moving targets and confusers; and (iii) a sensor resource manager that schedules camera pan, tilt, and zoom to support kinematic tracking, multiple target track association, scene context modeling, confirmatory identification, and collateral damage avoidance. When complete, VIVID MTT will enable precision tracking of the maximum number of targets permitted by sensor capabilities and by target behavior. This paper describes many of the challenges faced by the developers of the VIVID MTT component, and the solutions that are currently being implemented.*

Keywords: Video Tracking, Multiple Target Tracking, Sensor Resource Management, Air-to-ground tracking, UAV.

1 Introduction

Predator and other Unmanned Aerial Vehicles (UAVs) are revolutionizing targeting, carrying high performance Electro-Optical (EO) and Infra Red (IR) sensors to provide high quality data for target identification and engagement. Current targeting approaches using EO/IR sensors on UAVs depend completely upon human control of the sensor and interpretation of sensor data, effectively restricting these systems to supporting a single engagement at a time. The VIdeo Verification of IDentity (VIVID) program sponsored by the US Defense Advanced Research Projects Agency (DARPA) is addressing this problem by focusing on three key technology areas that include Collateral Damage Avoidance (CDA), Confirmatory Identification (CID), and Multiple Target Tracking (MTT). The program will conduct a series of critical technology and system demonstrations to create and refine automated target identification and verification technologies from both EO and IR motion imagery. The CID module creates high fidelity models of the targets to disambiguate confusion that may arise during tracking and to reconfirm a target's identity. The CDA module searches the vicinity of a designated target to detect people and vehicles that are not intended as targets. Alphatech is developing the VIVID MTT component (Fig. 1), which controls the sensor gimbal and interprets the video frames to maintain track of several distinct vehicles simultaneously—even when those vehicles are not all within the field of view (FOV) at the same time. MTT maintains track on multiple targets by slewing the sensor rapidly from target to target, and collecting just a few frames on each observation. Those frames can be used for updating the target state vector and continuing the track. The matching procedure embodied in the CID module can be used when necessary to disambiguate confusers that arise during tracking. The key objective of MTT is to maintain track of targets even in difficult conditions involving dense traffic and frequent occlusions. This track will also be used for real-time servo-control of a laser designator during the terminal phase of the weapon's flight.

Fig. 2 shows the main components of the MTT system, which comprises a front-end Video Processor (VP), a Multiple Hypothesis Tracker (MHT), and a Sensor Resource Manager (SRM) which interacts with the CID and CDA components of the VIVID system.

* Approved for Public Release, Distribution Unlimited.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **JUL 2004**	2. REPORT TYPE	3. DATES COVERED **00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE **A Multiple-Hypothesis Tracking of Multiple Ground Targets from Aerial Video with Dynamic Sensor Control**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defense Advanced Research Projects Agency (DARPA),Information Exploitation Office,3701 N Fairfax Drive,Arlington,VA,22203**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
7th International Conference on Information Fusion, 28 Jun ? 1 Jul 2004, Stockholm, Sweden.

14. ABSTRACT
The goal of the DARPA Video Verification of Identity (VIVID) program is to develop an automated videobased ground targeting system for unmanned aerial vehicles that significantly improves operator combat efficiency and effectiveness while minimizing collateral damage. One of the key components of VIVID is the Multiple Target Tracker (MTT) whose main function is to track many ground targets simultaneously by slewing the video sensor from target to target and zooming in and out as necessary. The MTT comprises three modules: (i) a video processor that performs moving object detection, feature extraction, and site modeling; (ii) a multiple hypothesis tracker that processes extracted video reports (e.g. positions, velocities, features) to generate tracks of currently and previously moving targets and confusers; and (iii) a sensor resource manager that schedules camera pan, tilt, and zoom to support kinematic tracking, multiple target track association scene context modeling, confirmatory identification, and collateral damage avoidance. When complete, VIVID MTT will enable precision tracking of the maximum number of targets permitted by sensor capabilities and by target behavior. This paper describes many of the challenges faced by the developers of the VIVID MTT component, and the solutions that are currently being implemented.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

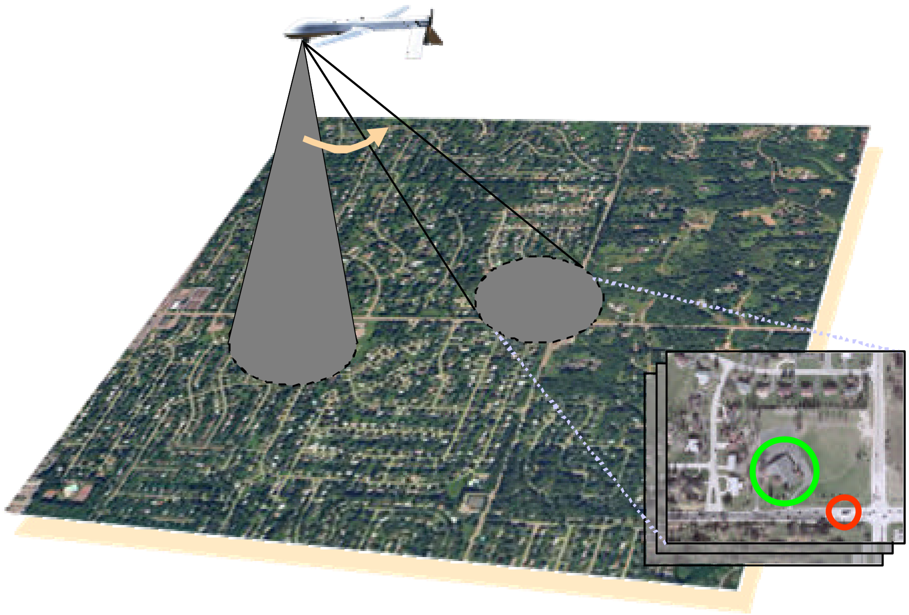

Fig. 1: The VIVID MTT system tracks multiple ground targets simultaneously by rapidly slewing the video sensor from target to target.

The inputs, outputs, and main functions of MTT are as follows.

- Inputs: (1) EO/IR imagery and current pan/tilt/zoom from the sensor; (2) target nominations and designations, areas of interest, and other commands from the operator; (3) reports from the CID module; (4) collateral damage assessments from the CDA module; and (5) NAV / IMU signals from the flight processor.
- Outputs: (1) gimbal and zoom commands to the Video Controller; (2) requests to the CID module (which include track ID and image chip history); (3) fire requests to the CDA module, and (4) target tracks, covariances, likelihoods, features, and image chips.
- Video Processor: The VP processes the incoming EO/IR imagery, performing frame-to-frame registration, motion segmentation, and feature extraction to detect and track moving objects and to create simple scene models. Registration and object detections/associations are fed to the MHT, which in turn feeds back filtered target position estimates, covariances, and predictions.
- Multiple Hypothesis Tracker: processes the line-of-sight (LOS) and feature reports from the Video Processor to create and update moving object tracks. Tracks comprise position and velocity estimates, feature estimates, error covariances, image chips (report history), and hypothesis likelihoods. Tracks are stored in the Track Database, which also contains the entire hypothesis tree and track predictions.
- Sensor Resource Manager: Based on the current nominated targets, confusers, and hypothesis tree status (uncertainties, ambiguities), the SRM generates sensor pan/tilt/zoom commands to optimize tracking performance metrics (number of high priority targets tracked, estimation errors, etc.).

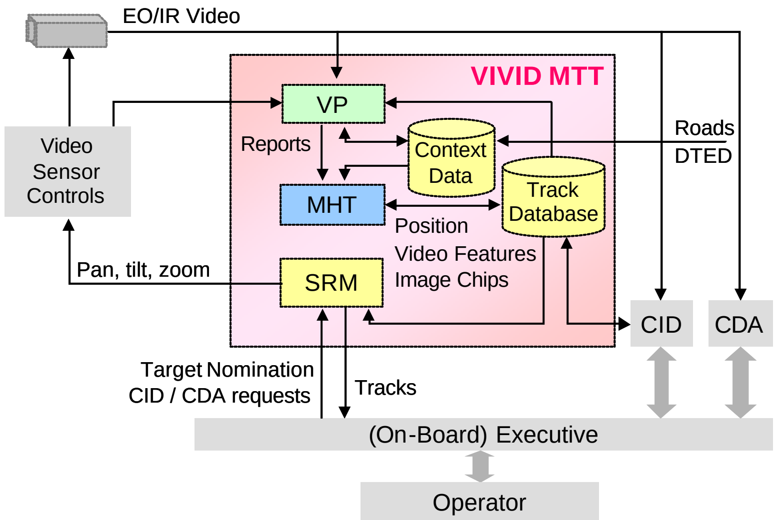

Fig. 2: The Multiple Target Tracker (MTT) processes incoming video stream, creates tracks, and generates FOV and slewing commands to track and identify several targets simultaneously. It also interfaces with the Confirmatory Identification (CID) and the Collateral Damage Avoidance (CDA) modules, which are the other main components of the VIVID system.

2 VIVID MTT concept of operation

The following representative tracking sequence illustrates the interplay between the three VIVID MTT components: VP, MHT, and SRM

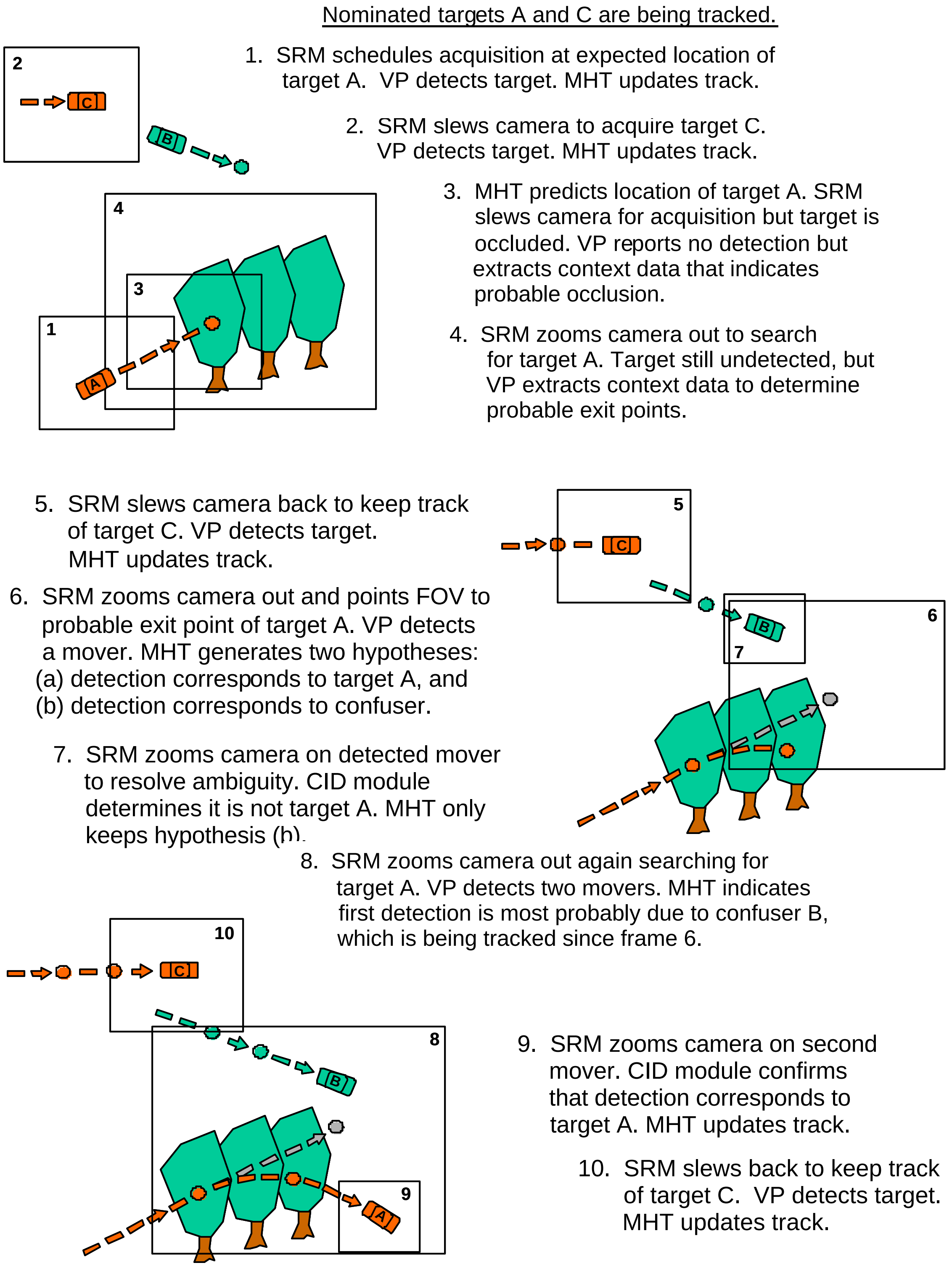

Fig.2: VIVID MTT optimizes pan/tilt/zoom/dwell time commands for tracking performance

3 VIVID MTT challenges

Some of the challenges that must be overcome by the MTT system become clear by examining a typical scenario in detail, as depicted in Fig. 4. Here, nominated targets (orange) are being tracked in the neighborhood of confusers (green). Since the targets are widely separated, the sensor must be slewed from target to target to maintain

track continuity and accuracy. During each sensor dwell, multiple image frames are collected at a zoom factor that permits the system to provide good detection probability, LOS accuracy, and vehicle features. The schedule of target contacts is chosen so as to minimize the slew time, and the commanded pointing angles come from predicted tracker estimates. The Inertial Measurement Unit's (IMU) estimates of these angles are enhanced by sequential video registration that is enabled via look-to-look FOV overlap. Furthermore, the FOV must be large enough to span the target location uncertainty that grows during the inter-revisit time (time to contact all targets). Solving this challenging problem requires tight interplay among advanced video processing, multi-target tracking, and sensor resource management technologies. Additionally, the system must optimize the target contact schedule, dwell times and zoom factors depending upon the continually changing target track environment.

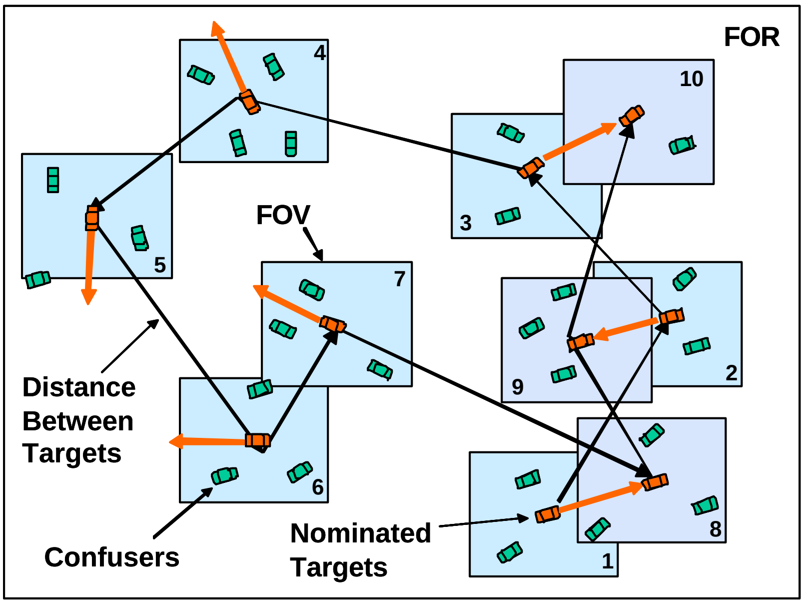

Fig. 2: The main objective of MTT is to track as many vehicles as possible with a single sensor.

As the camera slews from one point to the other, the video processor registers new images to previously-acquired images to create scene mosaics. Since these mosaics may not overlap, the VP maintains a dynamic list and determines which mosaic to operate on for each incoming frame. One situation that can arise is that the new image may overlap two or more existing mosaics. In that case, the video processor merges these mosaics into a single one. This requires the selection of a new common coordinate frame and a number of transformations to refer the old mosaics to the new coordinate frame. These transformations are propagated into the tracker, which needs to correct the tracks as well. The sensor resource manager accounts for all these issues to generate sensor commands.

4 System Components

As indicated in Fig. 2, the MTT system comprises three main components: a video processor, a multiple hypothesis tracker, and a sensor resource manager. A more detailed description of these components is provided in the following subsections.

4.1 Video Processor (VP)

The VP analyzes the video stream to support moving object detection, feature extraction, geo-location, and site modeling. Fig. 5 illustrates the high-level VP architecture and depicts the flow of raw imagery and meta-data from the sensor to key sub-components. A registration module compensates for camera motion by aligning individual video frames to a common reference using visual cues. Registration parameters are used to extract additional scene information, such as site mosaics and coarse 3D topology. A segmentation algorithm incorporates available site information, registered imagery, and guidance from the MHT tracker to determine the locations of targets and confusers. Spatio-temporal features for each object are also extracted to assist tracking, reducing reliance on potentially expensive CID requests where possible.

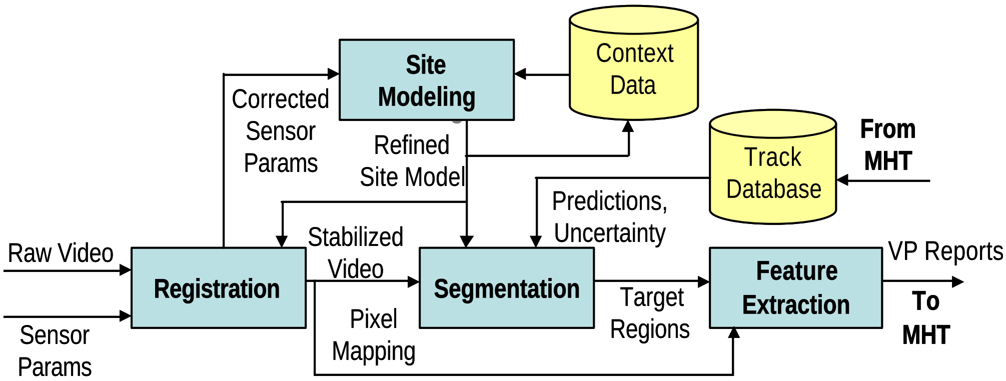

Fig. 3: The VP models the scene, registers imagery, and extracts target features.

4.1.1 Image-to-Image Registration

In order to provide contextual information to assist tracking, stabilize platform and camera motion, and correct metadata errors, the VP registers images to a common reference frame using a general plane-to-plane (8-parameter) perspective transformation model. Stabilizing the background essentially cancels camera motion, allowing targets to be tracked in a consistent, scene-relative coordinate space.

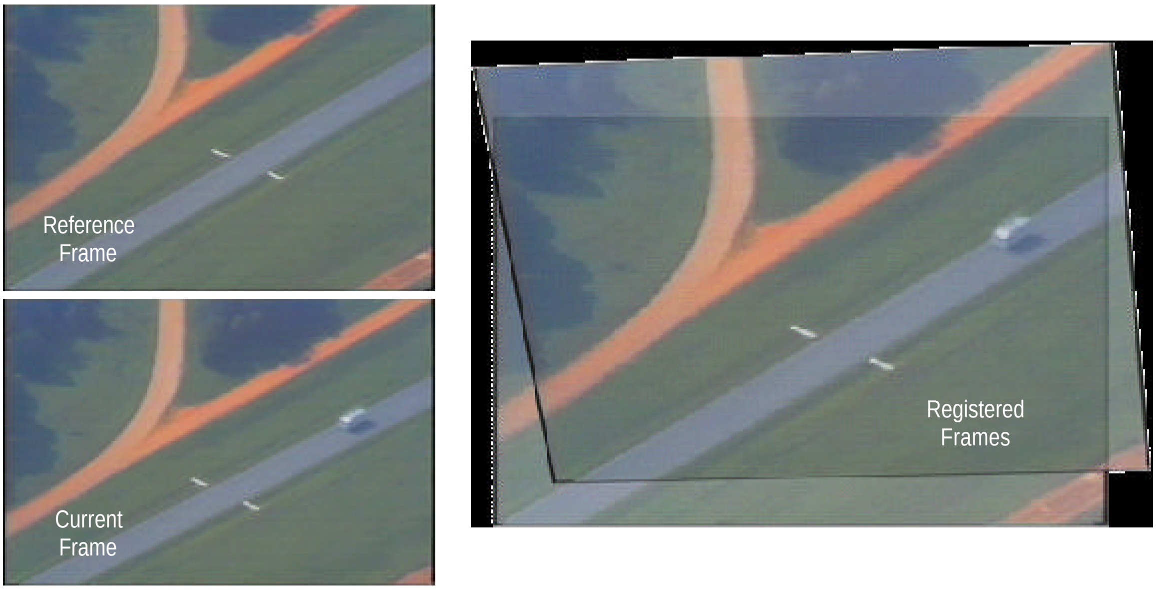

Fig. 4: Inter-frame registration aligns frames to a common reference.

Sequential registration, in which each frame is aligned with its immediate predecessor, requires minimal computation due to small inter-frame motion between adjacent frames. Motion between frame n and a “reference” (say frame 1) can be computed by concatenating these sequential transformations: . However, while registration is consistent in a local sense, small errors accumulate in the composite transformation so that frames n and 1 do not accurately align. To address this problem, we register each frame n directly to the reference frame, initializing with to assist algorithm convergence. Alignment becomes more difficult with higher n, since images become more and more dissimilar and effects such as parallax become significant with time; the VP therefore periodically updates its reference frame to a fresh image.

4.1.2 Motion Segmentation

With a set of images aligned on the background, the VP can extract motion by examining the registered color statistics at each pixel. To save computation, we recursively update aggregate statistics at each frame rather than examining all color values through a given spatio-temporal slice. The VP thus maintains a dynamic “background mosaic image,” which statistically encodes background pixel values; each new image is incorporated into the mosaic, and pixels that differ significantly from the background (in Mahalanobis distance) are flagged as “foreground,” grouped together into regions, and output to the MHT as reports in the current (registered) reference frame)

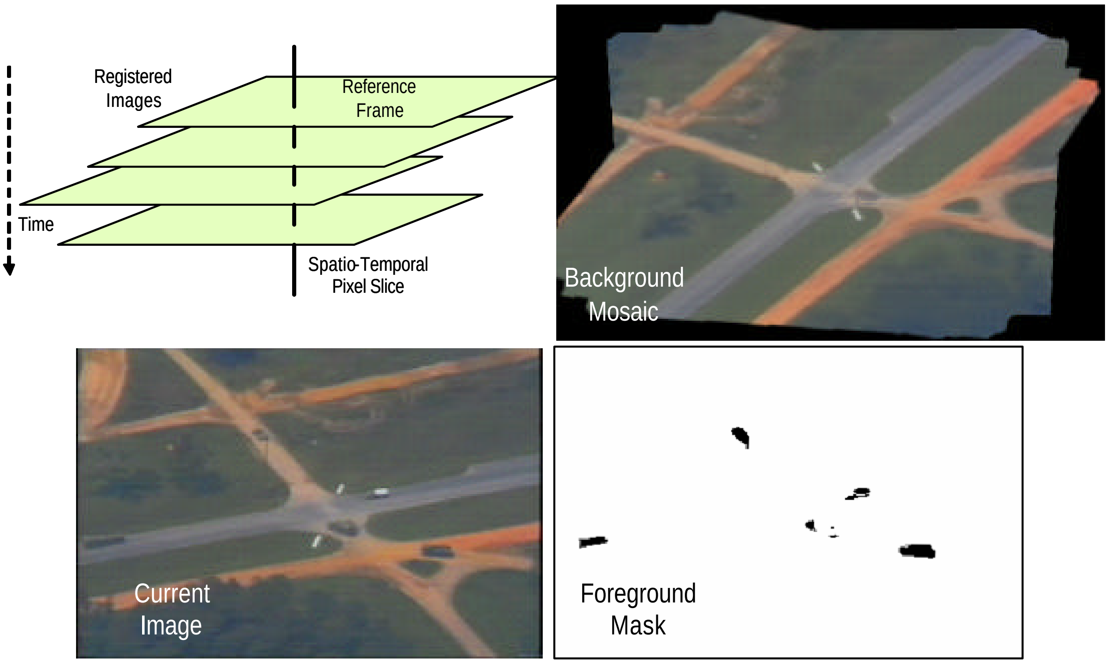

Fig. 5: Spatio-temporal statistical background segmentation (upper left) is applied to registered frames to form a background mosaic image (upper right). New images (lower left) are compared with the mosaic to form foreground mask containing moving objects.

4.1.3 Feature Extraction

The above method of object segmentation works well when frames can be registered—i.e., when the scene contains enough background texture to uniquely align the imagery—in alternate cases such as narrow FOV, different methods must be used to track a given target. The VP thus maintains an evolving set of simple features, or signatures, associated with each moving object that can be used to locate the object in new frames. Features are extracted and updated by the VP at each frame; several measurements, such as size and color, are also provided to the MHT, which maintains these features as part of each object's state in order to disambiguate targets at a coarse level without requiring CID queries

4.2 Multiple Hypothesis Tracker (MHT)

MHT technology is required to achieve VIVID goals because of high traffic densities and long revisit times due to camera slewing and prolonged target occlusion. ALPHATECH has developed MHT technology under a number of DARPA, AFRL, and ONR programs and is adapting and enhancing the tracker for this application. The MHT tracker works in coordination with the VP and SRM modules to maximize the performance of VIVID MTT system, using reports from the CID component when available. Fig. 8 shows a functional block diagram of the tracker with data flows to and from the VP and SRM modules, and from the CID component as well. The VP module sends moving object detections that comprise both object location and features such as object size and color. It also provides contextual data that is used by the tracker to improve target motion prediction. The SRM module supplies its near-future plans for camera slewing; these plans are used by the tracker to determine the optimum pruning policy. The CID component provides reports that are used by the tracker to resolve ambiguities. The tracker, in turn, processes all that information and generates tracks that comprise position and velocity estimates, feature estimates, error covariances, image chips (report history), and hypothesis likelihoods. These tracks are displayed to the operator and utilized by both the VP and SRM modules. The VP uses the tracks to improve scene segmentation, while the SRM module uses the current and predicted location of the targets to schedule the optimum slewing/acquisition strategy.

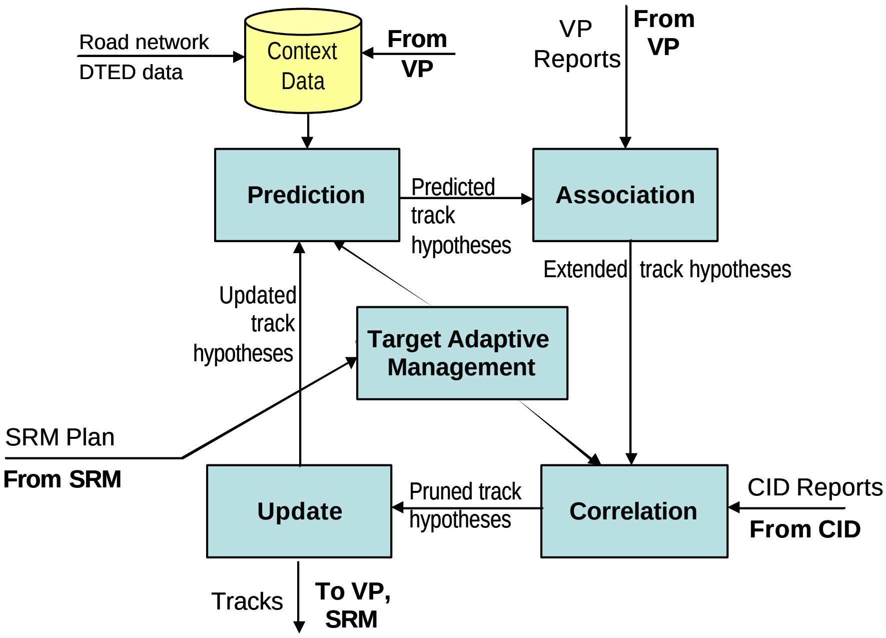

Fig. 6: The MHT tracker generates tracks and manages it own computational resources

4.3 Sensor Resource Manager (SRM)

The SRM manages the information collection necessary to support kinematic tracking, multi-target track association, confirmatory ID, context scene modeling, and collateral damage avoidance. It also tasks the laser designator for end-game weapon guidance. The SRM will balance the expected payoff of alternative viewing options against the costs due to sensor slewing, settling and collection time and task the sensor to optimize tracking performance. As a result, VIVID MTT will permit targeting the maximum number of targets consistent with the capabilities of the sensor and the behavior of the targets.

There are several technical challenges that must be overcome to apply an optimization-based approach to VIVID MTT. First, we must be able to quantitatively model the effect of sensor observations on tracking performance so that we can determine the relative value of alternative sensor viewing options. This requires determining such quantities as the probability of target detection, the number and quality of features extracted, and the accuracy of sensor observations as a function of such factors as target-sensor geometry, sensor mode (EO or IR), and zoom level. The impact of these quantities on tracking performance must then be modeled. Second, we must be able to predict the time to perform a sensor task as a function of these same factors. Due to sensor slewing, the time to perform a given sensor task depends on which previous task was performed. Both for purposes of valuing tasks and computing the time to perform them, we must estimate the position and possibly the orientation of targets relative to the sensor over the planning interval. This requires knowledge of the planned platform path and estimating the future positions of targets.

5 Visualization

Another challenge that needs to be overcome by the VIVID MTT system is the operator interface. For example, since the sensor slews very rapidly from target to target, spending only a fraction of second on each target or point of interest, the video stream can be unintelligible to a human operator. ALPHATECH is addressing this problem by developing a Video Visualizer that will stream the frames corresponding to distinct targets or points of interest to separate windows. Effectively, each separate window will look like the output of a sensor with a lower frame acquisition rate.

We have additionally leveraged existing visualization and interface tools developed under other programs. Fig. 11 shows a snapshot of the Display Manager (DM++) that we customized for VIVID MTT. This interface is used to display video reports and tracks, as well as other information that provides situational awareness to the operator, such as roads and maps. This interface is also used by the operator to nominate tracks and designate points of interest. The SRM uses that information to prioritize tracking of nominated targets over non-nominated ones, and to acquire imagery from the designated points of interest.

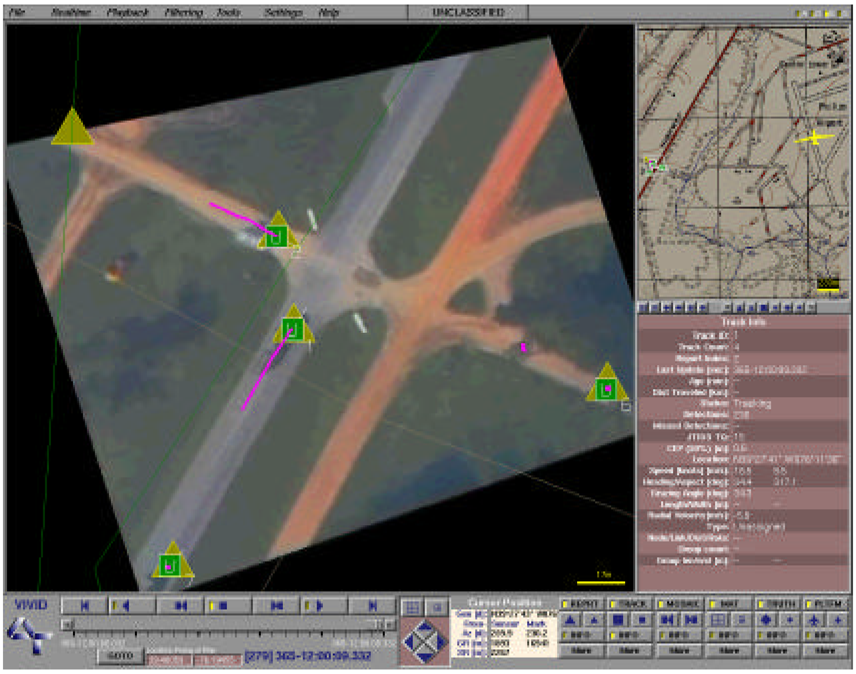

Fig. 7: DM++ tracking display and operator interface allows real-time visualization of situational information. The yellow triangles indicate video reports, the green squares indicate tracks, and the purple lines indicate track trails. Other information such as maps, mosaic overlays, and platform location is displayed as well.

6 Conclusions

In this paper we described some of the challenges that are being overcome for the development of the Multiple Target Tracking (MTT) component under the DARPA VIVID program. The main function of MTT is to track many ground targets simultaneously by slewing the video sensor from target to target and zooming in and out as necessary. Video processing, multiple hypothesis tracking, and sensor resource management technologies are being developed and tightly integrated to achieve the stringent performance requirements of the VIVID program. When the integration of the main three VIVID components—MTT, CID, and CDA—is completed, the system will exploit airborne optical and infrared video sources to track moving targets, confirm their identity, and to search the predicted impact area for collateral damage potential. This will improve precision targeting performance, limit friendly losses, and minimize collateral damage.

Acknowledgements

This work was supported by DARPA under contract # NBCHC030069.

References

[1] Stauffer, C. and W.E.L. Grimson. "Adaptive Background Mixture Models for Real-Time Tracking". In Proceedings of CVPR, 1999, pp. 246-252.

[2] Chao, A., and S. Berning, "Precise Image Calibration and Alignment," in Proceedings of the 55th Annual Meeting of the Institute of Navigation, June 1999.

[3] *Hartley, R., and A. Zisserman. Multiple View Geometry. Cambridge University Press, 2000.*

[4] Bertsekas, D. P. Dynamic Programming and Optimal Control, Athena Scientific, Belmont, MA, 2001.

[5] Washburn, R., M. Schneider, and J. Fox. "Stochastic Dynamic Programming Based Approaches to Sensor Resource Management." In Proceedings of 5th International Conference on Information Fusion, pages 608-615, 2002.

[6] Gittins, J. C., "Bandit Processes and Dynamic Allocation Indices", J. Roy. Stat. Soc., Vol. B, No. 41, 1979, pp. 148-164..

[7] Whittle, P., "Multi-Armed Bandits and the Gittins Index," J. Roy. Stat. Soc., Vo l. B, No. 42, 1982, pp. 143-149.

[8] Whittle, P., "Restless Bandits: Activity Allocation in a Changing World," Journal of Applied Probability, 25, 1988.